Noise reduction optimization of sound sensor based on a Conditional Generation Adversarial Network

Citation for the original published paper (version of record):
Lin, X., Yang, D., Mao, Y. et al (2021). Noise reduction optimization of sound sensor based on a Conditional Generation Adversarial
Network. Journal of Physics: Conference Series, 1873(1).
http://dx.doi.org/10.1088/1742-6596/1873/1/012034

Noise reduction optimization of sound sensor based on a Conditional Generation Adversarial Network

Xiongwei Lin[1], Dongru Yang[1], Yadong Mao[2], Lei Zhou[1], Xiaobo Zhao[1] and Shengguo Lu[1*]

[1] Guangdong Provincial Research Center on Smart Materials and Energy Conversion Devices, School of Materials and Energy, Guangdong University of Technology, Guangzhou, 510006, China

[2] School of Computer Science and Engineering, Chalmers University of Technology, Gothenburg 405 30, Sweden

[*] Corresponding author's e-mail: sglu@gdut.edu.cn

Abstract. To address the problems in the traditional speech signal noise elimination methods, such as the residual noise, poor real-time performance and narrow applications a new method is proposed to eliminate network voice noise based on deep learning of conditional generation adversarial network. In terms of the perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility measure (STOI) functions used as the loss function in the neural network, which were used as the loss function in the neural network, the flexibility of the whole network was optimized, and the training process of the model simplified. The experimental results indicate that, under the noisy environment, especially in a restaurant, the proposed noise reduction scheme improves the STOI score by 26.23% and PESQ score by 17.18%, respectively, compared with the traditional Wiener noise reduction algorithm. Therefore, the sound sensor's noise reduction scheme through our approach has achieved a remarkable noise reduction effect, more useful information transmission, and stronger practicability.

1. Introduction

With society's development and the improvement of informatization, voice signal as an essential communication carrier has become the primary human-computer interaction method[1]. However, the interference of various environmental noises makes the sound sensor cannot process the speech effectively in practical application [2]. It is necessary to preprocess the speech signal for optimizing the effect of speech processing in a noisy environment. Traditional speech de-noising algorithms include spectral subtraction, Wiener filtering and least mean square estimation[3-5]. Spectral subtraction is used to eliminate noise by subtracting unprocessed speech signal spectrum and noise signal spectrum, which has the characteristics of easy operation and high speed. However, rhythmic fluctuation "music noise" often appears in practical application, making the processing effect unsatisfied. The Wiener filtering method can minimize the difference between the expected signal and the mean square value of the actual output signal, which gets the optimal noise filter. However, this method is only suitable for the stationary noise signal. The minimum mean square error estimation (MMSE) de-noising method is a de-noising scheme based on statistical principle, which uses many parameter tests to determine the independent distribution of pure speech signal and noise signal. And the disadvantage of MMSE is that the real-time performance of signal processing is poor. The speech de-noising model based on deep learning can solve

traditional speech de-noising algorithms, determined by the characteristics of its deep nonlinear mapping network structure.

Its advantage is that it can approximate complex functions through a large number of training. Simultaneously, the neural network has the characteristics of multi-layers and wide width, which can be mapped to any function in theory and potential to solve complex problems. Besides, deep learning can also help input data-driven variables through the level by level learning method. Through the deep fitting of the data, we can avoid more accurate feature extraction. Therefore, the speech de-noising model based on deep learning has great potential. In 2013, Lu Xugang et al. proposed a greedy level by level pre-training strategy for training deep autoencoder (DAE), which significantly improved the performance of DAE in recovering pure target speech signal from noisy input speech signal[6]. In 2014, Xu Yong et al. proposed a speech enhancement algorithm using deep neural networks (DNN) with multi-layer depth architecture, enabling DNN to successfully separate pure speech signal from background noise without boring music pseudo-noise in a traditional noise reduction algorithm[7]. In 2015, Weininger et al. applied long short-term memory (LSTM) to speech noise reduction, which effectively utilized the characteristics of LSTM and achieved better performance in speech noise reduction [8]. In 2016, Balduzzi et al. proposed a speech noise reduction model based on recurrent neural networks (RNN) because of the time-series characteristics of speech signals, which significantly improved compared with DNN's speech noise reduction model [9]. In recent years, most of the speech de-noising models based on the traditional neural network are optimized by supervised learning of paired noisy and pure speech signals, and by iterative calculation to minimize the differences between the denoised pure speech signal. In this way, the distance function is usually based on Minkowski distance, which cannot sufficiently reflect human beings' subjective feelings for the processed speech signal. It can often not guarantee that the speech signal after noise reduction can achieve high quality and clarity. Therefore, we need a model that can autonomously learn the signal distribution mapping law to achieve better speech noise reduction according to more detailed speech evaluation index details.

Generative adversarial networks (GAN) is a new kind of deep learning model with self-game characteristics and self-optimization. It can deal with voice signals with strong correlation, strong time-varying and large data amount. In an ideal situation, speech signal follows Gaussian mixture distribution, and human speech signal and noise signal also follow Gaussian mixture distribution with different parameters [10]. By generating the self-game and self-optimization of GAN, the mapping relationship between pure speech and noisy speech can be well fitted to eliminate noise. To sum up, our work applies deep learning to sound sensors and proposes a condition generative adversarial networks (CGAN) speech de-noising method based on condition discrimination. Through the powerful data feature processing ability of the neural network, faster fitting speed and the self-game and self-optimization characteristics of GAN, the noise signal characteristics can be quickly analyzed from the noisy speech signal to achieve noise reduction. The noise reduction principle of the sound sensor is shown in Fig. 1.

2. The basic principle and improvement of Generative adversarial networks

2.1. The basic principle of Generative adversarial networks

Generative adversarial networks technology was first proposed by Goodfellow et al. The model consists of a generator (G) and a discriminator (D), and its core theory is the zero-sum game [11]. In GAN, the generator generates a new data sample by learning a large number of real data samples, and the discriminator is responsible for judging the authenticity of the data samples, which is equivalent to a classifier. The discriminator receives data samples and discriminates and then feeds back the corresponding information to optimize the generator, making the generator's final data closer to the real data. With the generator's continuous optimization, the generated data samples are closer to the real data samples, promoting the discriminator's ability to identify the authenticity. The two factors promote interoperability and finally achieve the dynamic balance. Fig. 2 is the flowchart of the GAN algorithm.

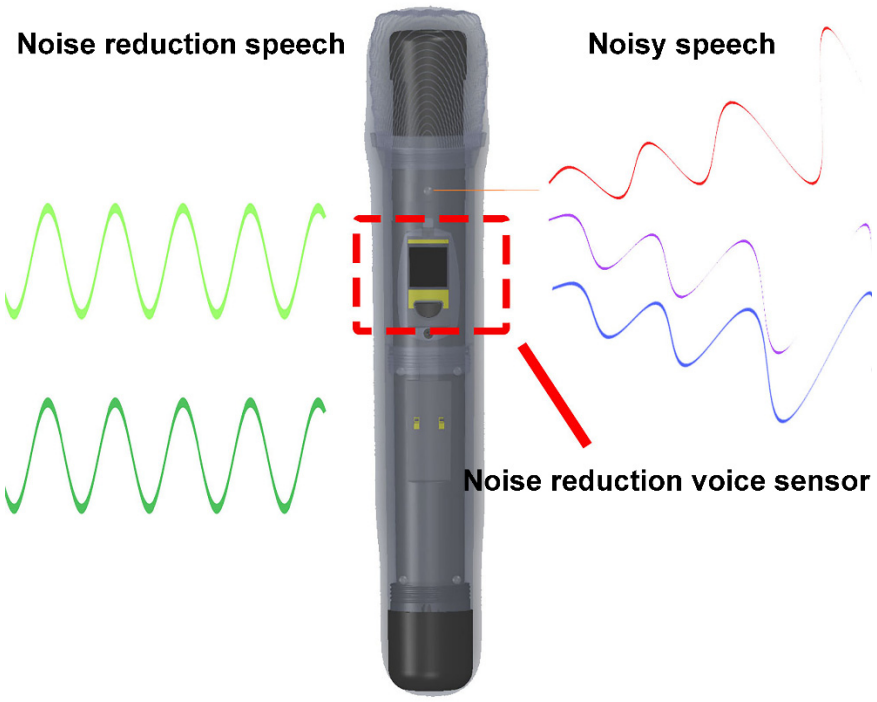

Figure 1. Schematic diagram of noise reduction principle for sound sensors.

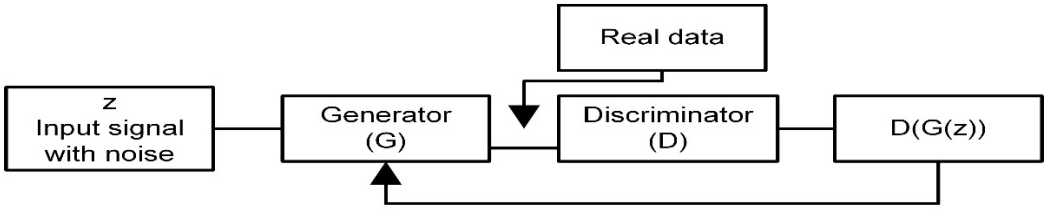

Figure 2. Flowchart of GAN's algorithm.

It is assumed that the pure voice data samples in GAN follow P_{data}'s $\{X_i\}_{i=1}^{N}$. P_g is the data sample generated by the generator. We use $P_g(x;\theta_g)$ to describe its data distribution. Input signal z with noise, then the neural network is used to approximate x :

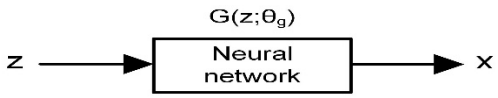

Figure 3. Flowchart of generator's algorithm.

The objective function of the generator can be calculated by the following formula (1):

$$G^{*}=\arg\min_{G}\mathbb{E}[\log(1-D(G(z)))] \quad (1)$$

The discriminator further classifies the data through neural network $D(x;\theta_d)$, when the value of $D(x;\theta_d)$ approaches 1, the higher the probability of input sample x coming from the real data sample; the closer $D(x;\theta_d)$ is to 0, the more approximate the input sample x is from the generator. The flowchart of the discriminator algorithm is shown in Fig. 4.

The objective function of the discriminator can be calculated by the following formula (2):

$$D^{*}=\arg\max_{D}\{\mathbb{E}[\log D(x)]+\mathbb{E}[\log(1-D(G(z)))]\} \quad (2)$$

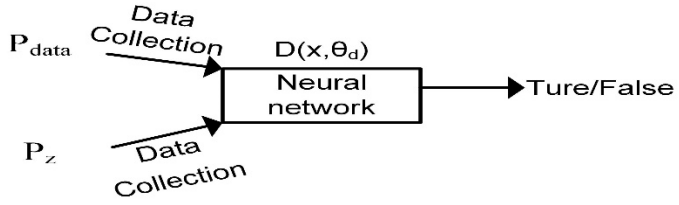

Figure 4. Flowchart of discriminator algorithm.

Combining formula (1) and formula (2), the total objective function V(D, G) is:

$$V(D,G)=\mathbb{E}_{x\sim P_{data}}[\log D(x)]+\mathbb{E}_{x\sim P_g}[\log(1-D(x))] \quad (3)$$

In the training process of generating a confrontation network, the generator's parameters need to be fixed first and optimize the parameter of discriminator D^{*}_{G}. Then fix the discriminator parameter D^{*}_{G} to get the optimal parameters of the generator:

$$G_{D^{*}}=\arg\min_{G}[\mathbb{E}_{x\sim P_{data}}(\log D^{*}_{G})+\mathbb{E}_{x\sim P_g}(\log D^{*}_{G})] \quad (4)$$

According to Kullback Leibler divergence theorem [12], , formula (4) can be transformed into formula (5):

$$G_{D^{*}}=\arg\min_{G}[KL(P_{data}\,||\,\frac{P_{data}+P_g}{2})+KL(P_g\,||\,\frac{P_{data}+P_g}{2})-\log 4] \quad (5)$$

The optimal solution is:

$$D^{*}_{G}=\frac{1}{2}\quad (P^{*}_{g}=P_{data}) \quad (6)$$

In formula (6), P^{*}_{g} is the data sample generated when the generator reaches the ideal level. Through the above operating, the generator and discriminator in GAN can achieve the optimal balance state and produce the best noise reduction effect.

2.2. Improvement of generative adversarial networks

However, GAN's discriminator makes a simple 0 / 1 evaluation with input data samples in practical application, which is too simple to generate better data samples. It may make the whole noise reduction model chain collapse in the training process[13]. To solve these problems, we use two speech signal quality evaluation functions (Perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI)) in the loss function for optimizing the training process.

2.2.1. Speech quality evaluation criteria. (1) Perceptual evaluation of speech quality(PESQ) speech quality perception evaluation is a standard method to evaluate speech quality[14], The principle of PESQ is aligning the pure speech signal with the test signal of PESQ in time and realize auditory transformation, and finally evaluate the quality of output speech through formula (7):

$$PESQ = 4.5-0.1\ d_{SYM}\ - 0.0309\ d_{ASYM} \quad (7)$$

Where d_{SYM} is the intensity of symmetric interference signal and d_{ASYM} is the intensity of asymmetric interference signal. Experiments show that the score range of PESQ is [0.5, 4.5]. The higher score with the higher speech quality and vice versa.

(2) Short-time objective intelligibility (STOI) is a new criterion for evaluating speech intelligibility proposed by Taal et al. in 2011[15]. Unlike other test models, which rely on global information to evaluate speech quality, STOI is an objective evaluation algorithm based on short-time (386 ms) signals.

The score range calculated by STOI is [0, 1]. The higher score with the higher speech quality and vice versa.

2.2.2. Improvement of the loss function. It is necessary to map the original signal x to the corresponding pure voice signal y through the minimum loss function in the generative adversarial networks model, as shown in the following formula (8):

$$L_{G(CGAN)} = \mathbb{E}_x \left[\lambda(D(G(x),x)-1)^2 \right] + \| G(x)-y \|_1 \quad (8)$$

In our research, we optimize the training mode of the discriminator firstly. For the loss function of the discriminator, the evaluation function PESQ and STOI are introduced, and the optimized discriminator function is shown in the following formula (9):

$$L_{D(CGAN)} = \mathbb{E}_{x,y} \left[(D(y,y)-Q(y,y))^2 + (D(G(x),y)-Q(G(x),y))^2 \right] \quad (9)$$

After optimization, the discriminator's evaluation for the received speech signal is no longer 0 or 1. It is based on the score obtained by PESQ and STOI, which can further evaluate the authenticity degree of the signal (the quality of the speech signal after noise reduction) and achieve a better discrimination effect and guide the machine's training more accurately. However, as the effective value range of PESQ is [0.5, 4.5] and STOI is [0, 1], it is necessary to weigh them, as shown in the following formula (10)

$$S(PESQ, STOI) = \alpha \times PESQ + \beta \times STOI \quad (10)$$

Where α is the weighting parameter of PESQ, β is the weighting parameter of STOI, and S (PESQ, STOI) is the score of the weighted speech signal, it can be seen from formula (9) that the loss function of a generator in CGAN has more advantages than that in traditional GAN, and the gradient convergence more effective. Therefore, the formula is improved as shown in formula (11):

$$L_{G(S-GAN)} = \mathbb{E}_x \left[(D(G(x),y)-S(PESQ, STOI))^2 \right] \quad (11)$$

The improved GAN model is a CGAN model with condition judgment.

3. Speech signal noise elimination method based on CGAN

3.1. The model of CGAN

The speech signal is a one-dimensional time-series signal, and the correlation between the sampling points of each signal is robust. Therefore, we can use bidirectional long short term memory (BLSTM) to process the speech signal. The model generator network in our conditional discriminant generative adversarial networks comprises one BLSTM level and two fully connected levels. Each level is composed of 200 nodes, 300 Leaky-ReLU nodes and 257 sigmoid nodes. The schematic diagram of the generator network structure is shown in Fig. 5.

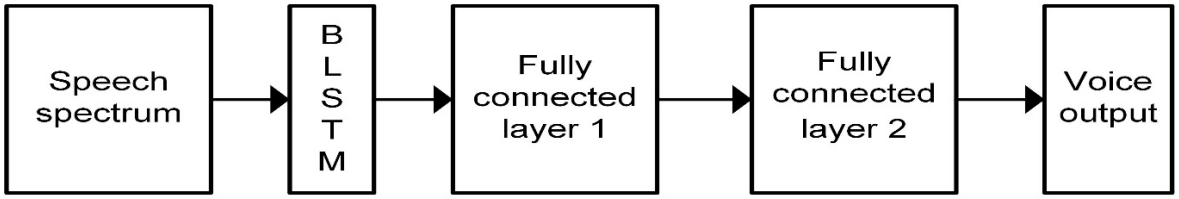

Figure 5. Schematic diagram of generator network structure.

The discriminator consists of five two-dimensional convolution levels, one two-bit global average pooling layer and three fully connected layers. The number and size of convolution kernels are: [10, (3,3)], [15, (5,5)], [25, (7,7)], [40, (9,9)], [50, (11,11)][16]. Through the two-dimensional average pooling layer, different lengths of speech input are fixed in 50 dimensions. The three fully connected levels are composed of 50 Leaky-ReLU nodes and 1 linear node—the Schematic diagram of the discriminator network structure in Fig. 6.

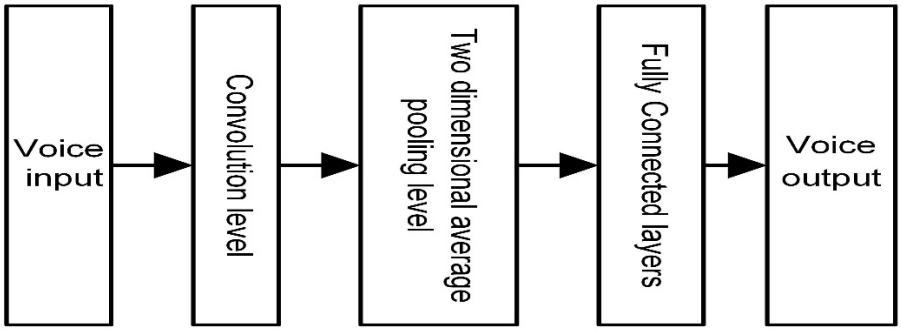

Figure 6. Schematic diagram of the discriminator network structure.

3.2. The Noise elimination model based on CGAN

3.2.1. Data preprocessing. Before the network model processes the voice signal, it is necessary to preprocess the input data. The preprocessing process includes pre-emphasis processing[17] and sub-frame plus windows processing[18]. We use FIR high pass filter (Finite Impulse Response) to enhance the high-frequency part of the signal at the beginning of the transmission line[19], to reduce the attenuation of high-frequency components in the transmission process and realize the pre-emphasis of the signal. Because the speech signal has short-term stability characteristics (it remains relatively unchanged in a short time range (10-30ms)), it can be used in frame processing. Because the speech signal has short-term stability characteristics (it remains relatively unchanged in a short time range (10-30ms)), it can be used in frame processing. After the frame length is 1024 and the frameshift length is 512, the jumping discontinuity is easy to appear between each frame's signals. To reduce the error, windowing is usually needed. We use the Hamming window[20] function to process the signal continuously. Finally, each frame's speech spectrum is obtained by short-time Fourier transform and input spectrum into GAN for training.

3.2.2. Training process. Firstly, the discriminator has initialized: the pure speech is input into the discriminator and the speech signal scoring function respectively to obtain the predicted score and the total score, and then the initial discriminator is obtained by the minimum mean square error method for the two scores. Furthermore, the generator is training: fix the parameters of the initialization discriminator at this time, input the noisy speech signal into the generator to get the denoised speech, and then get the predicted score through the discriminator; at the same time, the pure speech corresponding to the noisy speech gets the total score through the speech scoring function. The two scores are calculated by the least mean square error method, and then the parameters of the generator are optimized circularly until the noise reduction speech output by the generator can get the current maximum score.

The parameters of the generator are fixed to train the discriminator. The denoised speech and the corresponding pure speech generated by the noisy speech are input into the discriminator respectively to get the score. The discriminator parameters are updated by the minimum mean square error method so that the discriminator can make an accurate judgment. Repeat the above two steps until the generator's noise reduction speech can reach the maximum value of the speech scoring function. The whole GAN speech noise reduction model training process is shown in Fig. 7.

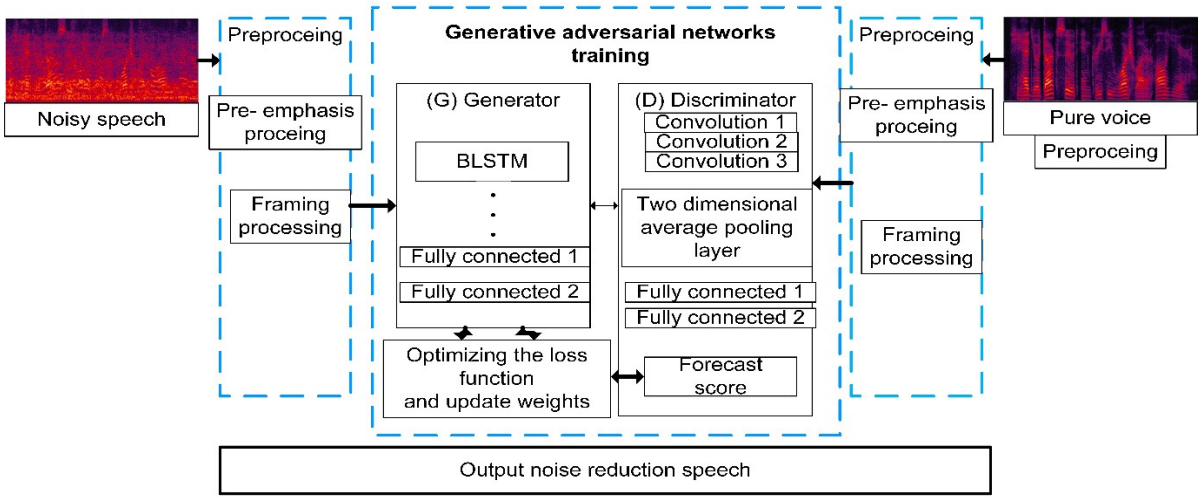

Figure 7. GAN speech noise reduction model training process.

4. Algorithm simulation experiment

To verify this method's actual noise reduction effect in the sound sensor chip, we recorded human audio in a quiet environment and noise environment, respectively. The noisy environment's background noise comes from three representative noises in the NOISEX-9 dataset[21]: restaurant environment, factory workshop and motor vehicle cockpit noisy environment. The recorded audio content is: "Welcome to China, you are welcome to Guangzhou!"

The verification speech of 5s duration was recorded in a quiet environment (≤ 30dB). The verification voice effect waveforms of 5s recorded in noisy restaurant environment after noise reduction is shown in Fig. 8; the verification voice effect waveforms of 5s recorded in noisy factory environment after noise reduction are shown in Fig. 9; the verification voice effect waveforms of 5s recorded in noisy motor vehicle cockpit environment after noise reduction are shown in Fig. 10. The contrast effect of speech spectrums before and after noise reduction in three environments is shown in Fig. 11.

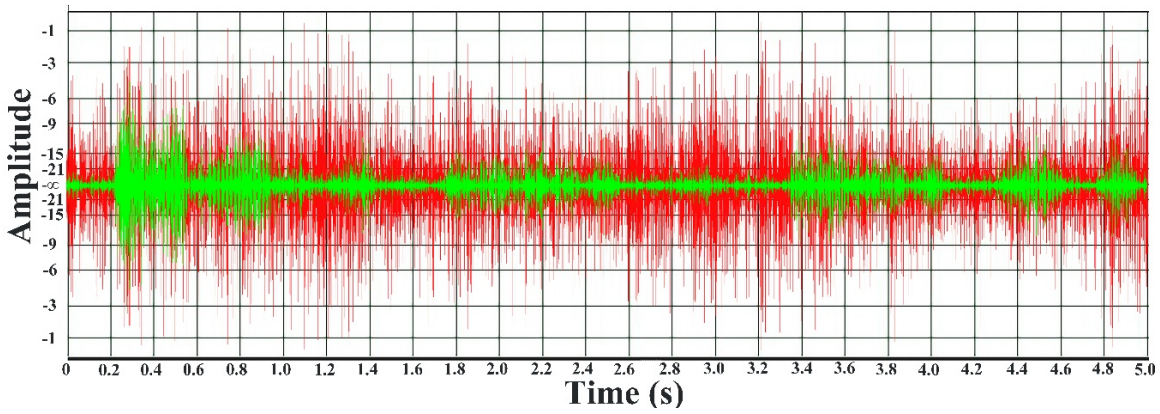

Figure 8. The waveform of voice noise reduction effect, which was recorded in a restaurant.

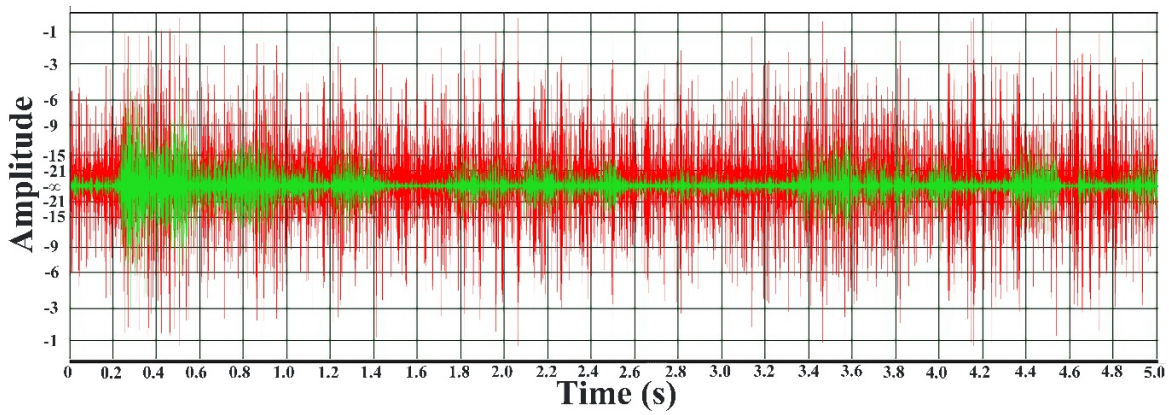

Figure 9. The waveform of voice noise reduction effect, which was recorded in a factory.

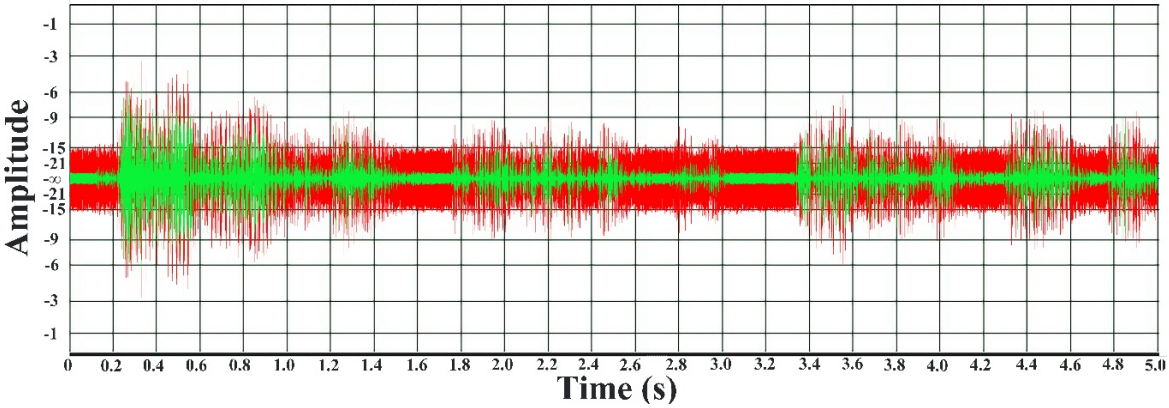

Figure 10. The waveform of voice noise reduction effect, which was recorded in the cockpit of a motor vehicle.

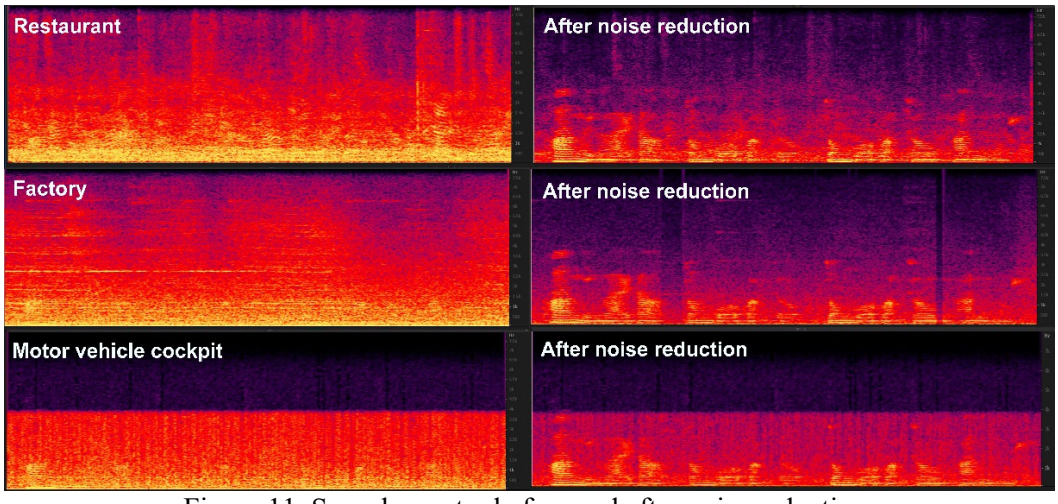

Figure 11. Speech spectra before and after noise reduction.

It can be seen that the noise reduction scheme for three kinds of daily noises has achieved a pronounced noise removal effect, which is close to the audio recorded in a quiet environment by comparing the waveforms and spectrum before and after noise reduction. The GAN method's noise reduction scores, classical spectral subtraction method, Wiener method, and MMSE filtering method are

compared by STOI and PESQ evaluation in the restaurant, factory workshop, and motor vehicle cockpit environment are shown in Fig. 7 and Fig. 8. In the three kinds of noise environment, the noise reduction scores of GAN are 0.77, 0.71 and 0.82, which are increased by 5.97 ~ 26.23 % compared with other noise reduction methods; by PESQ score, the three kinds of noise scores of GAN are 3.41, 3.33 and 3.49, which are increased by 6.11 ~ 17.18 % compared with other noise reduction methods. The results show that the STOI score and PESQ score of the algorithm are better than those of the traditional de-noising voice sensor, and the quality of the processed voice has been significantly improved.

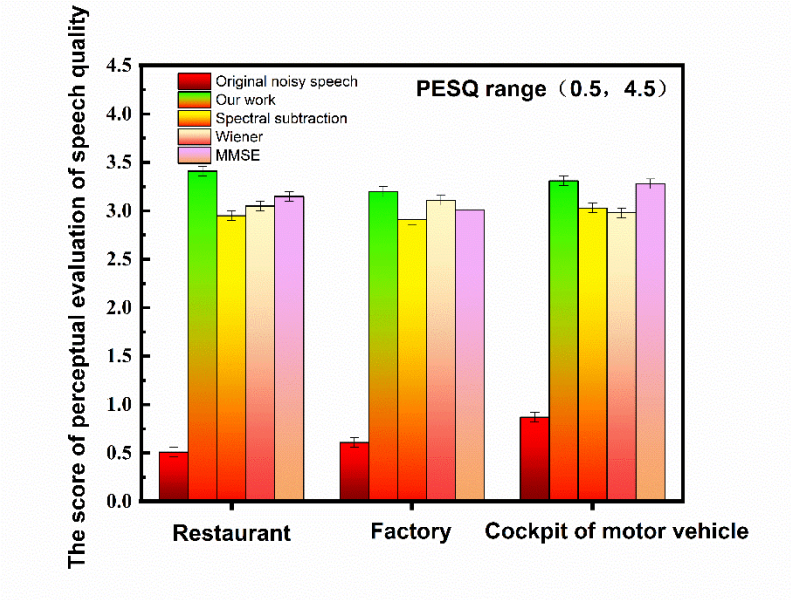

Figure 12. Comparison of PESQ evaluation scores.

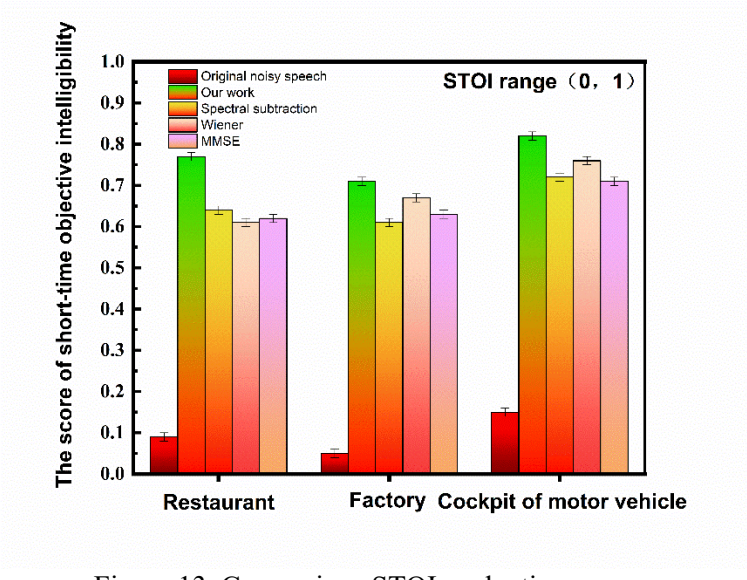

Figure 13. Comparison STOI evaluation scores.

5. Conclusion

In summary, the voice sensor based on the traditional noise reduction scheme is improved, and STOI score and PESQ score are introduced as the speech standard of condition evaluation generator to get better optimization system parameters. Using the structure of GAN and BLSTM, the generator can better extract speech features from the original noisy speech and pure speech and further generate better quality noise reduction speech. After conditional generation adversarial network processing noisy speech, the

voice sensor's noise reduction effect against the network voice signal is significantly improved than that of the traditional voice sensor. The sound sensor based on the generator's noise reduction model can be widely used in the front-end processing of speech recognition and other machines to improve speech recognition accuracy and speech processing speed. This method has a better future in obtaining higher quality speech in a noisy environment.

Acknowledgements
1.Project supported by the National Natural Science Foundation of China (Grant Nos. 51372042, 51872053)
2.The National Natural Science Foundation of China - Guangdong Joint Fund (Grant No. U1501246)
3.The Natural Science Foundation of Guangdong Province, China (Grant No. 2015A030308004)
4.The Frontier Research Project of Dongguan City, Guangdong, China (Grant No. 2019622101006)

References
[1] Gaikwad, S. K., Gawali, B. W., Yannawar, P. (2010) A review on speech recognition technique. Int. J. Comput. Appl., 10: 16-24.
[2] Shao, Y., Srinivasan, S., Jin, Z. (2008) A computational auditory scene analysis system for speech segregation and robust speech recognition. Comput. Speech Lang., 24: 148-151.
[3] Ephraim, Y., Malah, D. (1984) Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32: 1109-1121.
[4] Lim, J. S., Oppenheim, A. V. (1979) Enhancement and bandwidth compression of noisy speech. IEEE Proc., 67: 1586-1604.
[5] Boll, S. (1979) Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, 27: 113-120.
[6] Lu, X., Tsao, Y., Matsuda, S. (2013) Speech enhancement based on deep denoising autoencoder; In: proceedings of the Interspeech.pp. 436-440.
[7] Xu, Y., Du, J., Dai, L.-R. (2014) A regression approach to speech enhancement based on deep neural networks. IEEE-ACM T. on Audio, Spe., 23: 7-19.
[8] Weninger, F., Erdogan, H., Watanabe, S. (2015) Speech Enhancement with LSTM Recurrent Neural Networks and its Application to Noise-Robust ASR, Cham.pp. 91-99.
[9] Poroor, J. (2019) Low-level strongly typed dataframes for machine learning and statistical computing in resource-constrained devices; In: proceedings of the 2019 9th International Symposium on Embedded Computing and System Design (ISED).pp. 1-5.
[10] Pascual, S., Park, M., Serrà, J. (2018) Language and noise transfer in speech enhancement generative adversarial network; In: proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).pp. 5019-5023.
[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M. (2014) Generative adversarial nets; In: proceedings of the Advances in Neural Information Processing Systems.pp. 2672-2680.
[12] Kullback, S., Leibler, R. A. (1951) On Information and Sufficiency. Ann. Math. Statist., 22: 79-86.
[13] Salimans, T., Goodfellow, I., Zaremba, W. (2016) Improved techniques for training gans. 29: 2234-2242.
[14] Rix, A. W., Beerends, J. G., Hollier, M. P. (2001) Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs; In: proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221).pp. 749-752.
[15] Taal, C. H., Hendriks, R. C., Heusdens, R. (2011) An algorithm for intelligibility prediction of time–frequency weighted noisy speech. 19: 2125-2136.
[16] Pandey, A., Wang, D. J. I. a. T. O. A., Speech,, Processing, L. (2019) A new framework for CNN-based speech enhancement in the time domain. 27: 1179-1188.

[17] Liu, J., Lin, X. J. I. C., Magazine, S. (2004) Equalization in high-speed communication systems. 4: 4-17.
[18] Hamid, O. K. (2018) Frame blocking and windowing speech signal. Journal of Information, Communication, Intelligence Systems, 4: 87-94.
[19] Moulines, E., Duhamel, P., Cardoso, J.-F. (1995) Subspace methods for the blind identification of multichannel FIR filters. 43: 516-525.
[20] Shahin, I. (2008) Speaker identification in the shouted environment using suprasegmental hidden Markov models. Signal Process., 88: 2700-2708.
[21] Varga, A., Steeneken, H. J. (1993) Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech communication, 12: 247-251.